Back to the Basics: Curriculum Reform and Student Learning in Tanzania

Daniel Rodriguez-Segura and Isaac Mbiti

Abstract

In 2015, the Tanzanian government implemented a curriculum reform that focused instruction in Grades 1 and 2 on the "3Rs"—reading, writing, and arithmetic. Consequently, almost 80 percent of the instructional time in these grades was mandated towards foundational literacy in Kiswahili and numeracy skills. Other subjects such as English were no longer taught. Using student-level panel data, we evaluate the effect of this policy on learning outcomes using a difference-in-differences approach which leverages the variation in the timing of implementation across grade levels and cohorts impacted by the policy. We find that the policy increased learning by around 0.20 standard deviations in Kiswahili and math test scores one year after the start of the reform. Timely teacher training on the new curriculum was associated with even larger effects. Evaluating longer term outcomes, we find suggestive evidence that the reform decreased the dropout rate of children up to four years later. However, this was also accompanied with lower average passing rates in the national Grade 4 examination due to compositional changes as low-performing students became less likely to dropout.

Back to the Basics: Curriculum Reform and Student Learning in Tanzania

Daniel Rodriguez-Segura
University of Virginia, School of Education and Human Development
dan.rodriguez@virginia.edu

Isaac Mbiti
University of Virginia, Frank Batten School of Leadership and Public Policy
imbiti@virginia.edu

Research on Improving Systems of Education (RISE)

www.riseprogramme.org

information@riseprogramme.org

I. Introduction

Curricula are a key input of any educational system. Ideally, an educational system's intended curriculum determines the material mandated to be taught in school and the desired instructional approaches. Yet in practice, curricula in many developing countries are often too expansive or "overambitious" relative to their education system's capacity (Pritchett, 2013). There are also concerns that these curricula favor children from advantaged backgrounds (Glewwe, Kremer, and Moulin, 2009). Because teachers have incentives to cover the entire syllabus, scholars have hypothesized that a wide curriculum could encourage teachers to either increase the pace of instruction beyond the rate of student learning, to focus their attention on the students who can keep up, or both (World Bank, 2017; Pritchett and Beatty, 2015; Muralidharan and Zielenkiak, 2014). This, in turn, could be a potential explanation of why the progression of student learning in many developing countries is slow, with very few students demonstrating appropriate grade-level competencies, and the majority of them having a mastery level several grades below where they should be (World Bank, 2018; Pritchett and Beatty, 2015).

The Tanzanian education system, which we study, exhibited many of the characteristics associated with overburdened education sectors for the past two decades. These included a curriculum featuring numerous subjects, low levels of learning relative to international benchmarks, slow learning progression, and trademarks of systems burdened by expansive curricula (Ministry of Education, 2015; USAID, 2015). For instance, prior to 2015, students in early primary school (Grades 1 through 3) were taught eight different subjects, including Information and Communications Technology (ICT) and agriculture. The learning profiles were flat with foundational numeracy and literacy skills gained slowly over time (Jones, Ruto, Schipper, and Rajani, 2014). Consequently, most students fell behind the prescribed curriculum– only 31 % of grade 3 students were proficient at the grade 2 level (Jones, Ruto, Schipper, and Rajani, 2014), and the majority of grade 4 students had not mastered grade 3 material (World Bank, 2017).

Faced with the growing evidence on the low learning levels in early grades, the Tanzanian government enacted the 3Rs reform (Reading, wRiting, and aRithmetic), also known as the "3Ks" in Kiswahili, for grades 1 and 2. This reform was enacted in the 2015 school year and narrowed the scope of the grade 1 and 2 curriculum such that 80% of the instructional time would focus on the 3Rs, with all literacy focused on Kiswahili rather than English as it had been previously. English, which was taught as a subject in the first two grades was removed from the curriculum and reintroduced starting at Grade 3. While proponents of narrow curricula argue such reforms are likely to benefit most learners, the

potential benefits might not materialize due to state capacity constraints, such as teacher training. Further, the potential benefits on numeracy and literacy may come at the expense of non-focal subjects, and the reforms may constrain the potential of high performing students who likely benefit from a faster pace. In fact, in wealthier contexts like the United States, "curriculum narrowing" often comes with negative connotations linked to the unintended consequences of test-based accountability and the excessive focus on a handful of tested subjects. The slower pace may also generate a compositional effect if students who would have fallen behind under the expansive regime are less likely to drop out because of the reform.

Despite the ubiquity of overambitious curricula in developing contexts, there is limited causal evidence on the potential for content reducing curriculum reforms to improve student learning outcomes. This is partly due to the challenges of credibly estimating the casual impact of a nationwide reform which affects all students simultaneously, and the lack of adequate data. The reforms that have been studied have focused evaluating different targeted instruction models (e.g., Banerjee et al., 2017, Muralidharan *et al*., 2019), and changes in the language of instruction (e.g., Ramachandran, 2017; Seid, 2019, and Laitin et. al., 2019). In this paper, we examine the consequences of this 2015 Tanzania curriculum reform using a unique student-level panel dataset of students from grades 1-3, drawn from a large nationally representative randomized control trial (Mbiti et al., 2019 and Mbiti et al. 2021). To estimate longer run effects, we use administrative data on national test scores in grades 4 and 7 (the last grade in primary school). This allows us to explore the effects of the reform on both passing rates in national examinations four years after it was implemented, as well as the relative change in the number of test-takers in grade 4 (which acts like a proxy for downstream enrollment.)

We identify the impact of the 3Rs reform using a difference-in-difference strategy that takes advantage of the variation in student exposure to the reform by grade level. Specifically, we compare test score outcomes among students in the first two grades (treated grades) to the test scores among third graders (comparison grade), pre- and post-reform (2014 compared to 2015). To explore longer term implications of the reform, we use the administrative data on grade 4 and 7 national test scores to estimate the effect of the program on learning and school enrollment four years after the reform was first implemented. In particular, we compare outcomes for pupils in grade 4 in 2018 —the cohort of students who was in grade 1 in 2015— to the outcomes of pupils in grade 7 that the same year, as grade 7 pupils did not experience the curriculum reform during the period covered by our data. Given that we have access to the universe of national test scores for these examinations, we also explore whether the number of test-takers increased as a result of the policy to proxy for school enrollment outcomes four years later.

We find that the nationwide curricular reform produced moderate average gains in numeracy and literacy by approximately 0.20 SD. These gains were equivalent to a reduction in pre-policy learning gaps between the top and bottom wealth quintiles of 28% in both math and in Kiswahili. Similarly, we can rule out negative effects on English, a subject de-emphasized by the reform, larger negative effects than -0.02 SD. We also find that learning grains were larger in schools that received timely teacher training on the new curriculum, providing suggestive evidence for the importance of proper implementation, especially for a governmental curricular reform of this scale. In the longer-term, the policy increased the number of students taking the fourth-grade national test by 16%, suggesting that the policy improved student retention and grade progression. The improvement in student grade progression was also accompanied by decreases in the passing rate of these examinations. However, these decreases were comparable in magnitude or smaller than what would be expected given the overall increase in the number of test takers, suggesting that there was still an increase in the aggregate level of learning in Tanzania as a result of the reform.

Our study makes three distinct contributions. First, it is one of the few studies that examines the causal impact of a narrower curriculum on learning outcomes in a developing country. Despite the recognition of overcrowded curriculums in developing countries (Atuhurra and Alinda, 2018; Atuhurra and Kaffenberger, 2020; Pritchett and Beatty, 2015), there is limited causal evidence on the potential impact of reducing the required instructional content. The literature that estimates the causal impact of curriculum reforms on learning has generally focused on the effects of (large-scale) changes to the language of instruction in schools from colonial languages of instruction to local languages (or mother tongue). The debate on language policy is less relevant for primary education in Tanzania because, unlike other countries in regions, the language of instruction in primary schools has been Kiswahili (rather than English) since the 1960s. Overall, the evidence on the effectiveness of reforming the language of instruction on student learning in mixed. For instance, Ramachandran (2017), Seid (2019), Laitin et. al (2019), Brunette et. al (2019) and Kerwin and Thornton (2020) find that that these reforms improve learning outcomes, whereas as Piper et. al (2018) and Chicoine (2019) show that such reforms fail to improve learning. Our study fills this gap by using a credible identification strategy, coupled with student panel data to show that such reforms can improve student learning in early grades. Further, we show that the learning improvements across almost all measured sub-domains of numeracy and literacy (for example, two-digit addition and word recognition).

Second, we use administrative data from 2015-2018 to examine the longer run impact of the reform, which has been broadly hard to quantify in the international education literature due to the lack of appropriate data. Our results show that the student's fully exposed to the reform were more likely to take the fourth-grade exam. This could reflect the

improved retention and grade progression effects of the reform. However, the differences-in-differences estimates on learning are negative, potentially reflecting the compositional change in the sample towards a lower-performing student body on average.

Third, we use our data to examine potential mechanisms. We focus on implementation – specifically the extent to rollout of teacher training on the curriculum. Developing nations often face challenges implementing policies, programs, and reforms at scale, muting their potential beneficial effects (Banerjee et. al, 2017; Bold et. al, 2018). Students in schools with at least one teacher trained in the 3R reforms had better learning gains compared to the counterparts in schools with no trained teachers, although this difference was not statistically significant.

Our work contributes to the literature on the potential for curriculum reforms that narrow the instructional content to improve learning outcomes when implemented in contexts in which the curriculum has previously been overcrowded or overly ambitious. These reforms are arguably extremely relevant for developing country contexts where instructional time is limited due to teacher absenteeism (World Bank, 2018) and the inclusion of multiple (potentially tangential) subjects can crowd out the teaching of core competencies such as the 3Rs. Even outside of primary and secondary education systems in developing countries, similar debates are ongoing regarding the potential deleterious effects of an overcrowded curriculum in medical schools in developed countries (Slavin and D'Eon, 2021). In this way, this study offers some initial evidence on the potential benefits, and unintended side effects, of such a reform.

II.   Context

National curricula often reflect the political priorities, historical roots, and sociocultural environment in which schools operate. For example, the curricula in many developing often contain features from past colonial institutions, such as retaining English or French as the language of instruction (Mwiria, 1991; Malisa and Missedja, 2019; Erling and Hultgren, 2017). On this dimension, Tanzania has been an exception relative to its neighbors – the language of instruction in primary schools has been, for the most part, Kiswahili rather than English since at least 2007 (Sa, 2007).

Primary school in Tanzania comprises seven grades. While the net enrollment rate in primary school increased from 53% in 2000 to 80% in 2014, in 2015 only 35% of third graders and 72% of grade 7 students met minimum learning benchmarks suitable for second graders (Twaweza, 2017). These low levels of learning are coupled with large geographic and socioeconomic disparities. For instance, the urban-rural gap in 2015 was about 0.5 standard deviations in test-based math and Kiswahili performance, while the gap between the top and bottom wealth quintiles was about 0.7 standard deviations (Twaweza, 2015).

As of 2013, the Tanzanian curriculum for grades 1-2 was an archetypical overambitious curriculum, consisting of eight subjects, including "Vocational Skills", "Information and Communication Technology", and "Personality"[1]. In the words of the Ministry of Education and Vocational Training,

> "The Curriculum for Standard I and II was overloaded with subjects, causing teachers to overemphasize the teaching of subject content and placing less emphasis on the development of the basic skills and competences in Reading, Writing and Arithmetic that are necessary in order for learners to effectively learn content." (Tanzanian Government Policy Report, 2016)

Twaweza, a well-known East African civil society organization, speaks directly to this issue and reports that, "the learning expectations implied by the curriculum are that children rapidly master basic reading skills in both English and Kiswahili, as well as basic numeracy skills up to multiplication[...] Contrary to curriculum expectations, the data show that many children in Tanzania do not master these basic skills quickly" (Twaweza, 2017). Furthermore, they highlight that although by the end of third grade students are expected to have mastered basic numeracy and literacy, students continue to develop these skills in later years (Twaweza, 2017). In sum, government agencies and external observers were in agreement that before the 3Rs reform, the curricular expectations and students' learning levels were clearly misaligned due to the presence of a typical overambitious curriculum.

III.   Policy reform

In response to the weak learning levels, the Government of Tanzania implemented the Big Results Now in Education (BRN) Initiative in 2013. The BRN policy included nine reforms, ranging from the mandated public release of within-district school rankings to infrastructure improvement to teacher training, all of which were rolled out at different times. The different policy changes are described in greater depth in Appendix D. In general, these policy changes did not overlap in terms of content or grades targeted by the reforms that we study here and therefore do not confound our main estimates. However, we still conduct empirical checks for potential heterogeneity based on these other reforms in Section V.

One of these policy changes, and the focus of this study, centered on curricular reform for grades 1-2, implemented in 2015. The aim of this reform was to strengthen the "3Rs" reading, writing, and arithmetic by allocating a larger share of the existing instructional time to numeracy and literacy. The new de jure allocation of time was such

---

[1] The inclusion of "technical" subjects from an early age appears to be a colonial legacy (Malisa and Missedja, 2019), as Mwiria (1991) also describes: "in keeping with a colonial ideology which stressed the role of the African as that of service to the white man, technical and agricultural (as opposed to academic) education were recommended for Africans by the missionaries, colonial authorities and external educational commissions".

that 80% of the instructional time was supposed to be spent working on the three core skills through the school subjects of math and Kiswahili. The remaining 20% of time was allocated for the other subjects. The school day was not extended, and therefore the policy entailed a re-allocation rather than an increase in class time. English was officially removed from the grade 1-2 curriculum, in an effort to focus on literacy in Kiswahili, Tanzania's national language. Under the revised 2015 curriculum, English is only taught starting in grade 3[2]. Finally, content from some of the other subjects that had been removed from the new curriculum was incorporated into the curriculum via Kiswahili reading passages on science or social studies.

In practice, the change in the time allocation for numeracy and literacy did not increase all the way to the mandated 80% of instructional time by 2015 – yet, the change was sizable, as we show in Table 1[3]. We use data from class observations and government documents such as the policy report in Tanzanian Ministry of Education (2015) to estimate that before the reform, roughly 45% to 60% of the time was devoted to the "3R", including English lessons[4]. Using similar class observation data from 2015, we place the lower bound of the increase in instructional time for 3Rs at 1.3 hours, or ~14% from a base of 9.2 hours in the observational data. In other words, after the reforms, 70% of the total instructional time was devoted to the 3Rs, on average. This estimate includes English lessons, which should have technically not been taught post reform but that we still detect in our observational data. When we consider only math and Kiswahili, the focus of the policy, the increase in instructional time in total for both subjects in grades 1-2 was 2.4 hours, or 39% from a base of 6.2 hours per week in the observational data (that is, 57% of the total instructional time would have been devoted to the 3Rs, as opposed to the mandated 80%). In turn, this increase of 39% in instructional time towards the 3Rs serves as our upper estimate of the effect (i.e., the "first stage") of the policy in practice. The midpoint between these two bounds is an increase of 1.9 hours per week during a week that expects 15 hours of instruction. In other words, our mid-range estimate is that 12 additional percentage points, or almost two additional hours per week, were devoted to the 3Rs as a result of the reform.

An important factor to understand how this particular reform "slowed down" the pace of the curriculum is to understand which curricular inputs changed. In other words, if

---

[2] While teachers indeed reported following this change post-reform in observational and survey data, in practice, we observe that this change did not happen as suddenly, and while the observed time of English instruction decreased ~40% post-reform, it was still being taught for about 2 hours per week (see Figure 1).

[3] The numbers displayed on this figure come from observational data collected for the original studies, only available for 2014 and 2015. While this data displays a large degree of missingness, especially in 2014, we were not able to find any reliable source quantifying precisely the allocation of time for the core subjects pre-reform.

[4] This is in rough agreement with the 9.2 hours (61% of the time) that we quantify on average for grades 1-2 for 2014 in the observational data.

the pace of curricula is defined as "content covered" over "time allocated to this content", the slowing down of the pace of an overambitious curriculum could happen through a decrease in the amount of content which is expected to be covered in class, an increase in the time allocated to this material, or both. Although the policy documents and curriculum descriptions do not explicitly mention which avenue was pursued by the Tanzanian government, we do not find any evidence that the expected amount of content to be covered within the "3Rs" changed in any way (World Bank, 2015, 2016, 2017; Ministry of Education, Science and Technology, 2016). Instead, from our teacher observation data in 2014 and 2015 and the main official policy description (Ministry of Education, Science and Technology, 2016), it seems like this reform slowed down the pace of the curriculum (for numeracy and literacy) almost entirely through the channel of time re-allocation towards these subjects.

The spirit of the policy reform was aligned with best practices to improve learning levels that researchers and donor institutions have advocated for, and that have been effective in other interventions which have sought to better align instruction with student learning levels like "Teach at the Right Level" (for instance, in Banerjee et al., 2017). However, there are several reasons why it is not a certainty that learning would increase after implementing curricular reform at a national level. First of all, as expected in contexts with weaker state capacity, the implementation of the policy was not standardized, and not all teachers and schools received the same materials and degree of government support (Komba and Shukia, 2021). For instance, while 93% of schools claimed to have changed the curriculum to 3R in 2015, 4% of these also claimed to still teach English and Kiswahili in the grade 2, something that was explicitly contrary to the policy[5]. Similarly, not all teachers received the training on time: only 37% of all the teachers in our sample received the training, and 96% of these got it in 2015, after the school year had started. The distribution of materials was similarly scattered: from our survey data, we estimate that 4 of every 10 teachers in our sample do not have any textbooks that reflected the 3R curricular changes, and even among those that do, the kind of materials varied. Half of the teachers with textbooks that reflected the 3R curricular reform had them for writing and math, but only one third had books for reading[6]. Secondly, even when the underlying mastery of skills of socioeconomically disadvantaged children in LMIC is improved through educational

---

[5] Still, the rollout of the teacher training was gradual, but it ended up covering almost all schools: in 2014 only 14% of all schools had at least one teacher that had received training on the 3R curriculum, but by 2015, 99% of schools claimed to have at least one teacher who had received the training.

[6] In spite of all these logistical challenges, according to our survey data, 72% of all head teachers think that the implementation of the 3R reform went "well" or "very well". Note that most of these figures were calculated using the full experimental sample, not just the control schools, as it is the case with most of the other figures in this paper (see the Data section for more details on the sample). Tests of statistical significance showed no systematic differences between these two groups, so we chose to use the full sample to increase the precision of these numbers.

interventions, work such as Dillon et al. (2017) shows that these gains may not translate into gains in formal test scores, displaying the potential gap between not only curriculum and children's knowledge, but also children' knowledge and performance on assessments. Because of these reasons, the study of this particular reform is valuable to begin to understand whether the best practices of curricular alignment can indeed be implemented at scale by mostly government entities, and eventually be reflected in traditional learning measurements.

IV. Research design

Main learning panel

The main data source for this project was collected through the KiuFunza I and KiuFunza II projects, conducted in Tanzania between 2013 and 2016. These projects were randomized controlled trials studying school incentives and teacher bonuses respectively (Mbiti et al. 2019; Mbiti et al., 2021). Both studies included a core set of 180 schools from 10 districts. For this study, we focus on students in the 60 randomly selected schools which served as the control schools for the original RCT studies. Since the original experimental sample of Mbiti et al. (2019) consisted of a set of nationally representative public schools, and the current paper uses a subset of randomly-selected schools from this sample (that is, the control group from the RCTs), the current sample also consists of a sub-sample of nationally representative schools. It is worth noting that this control group did not receive any of the incentives that treatment schools did, and served solely to benchmark the effects of the other interventions. In other words, these schools would have been exposed only to the same policy and input changes as all other schools in Tanzania over this period[7].

Within these 60 schools, we have a longitudinal panel of grade 1-3 students for three years, 2014-2016[8] (although our main specification uses 2014-2015 for reasons that will be detailed in the next section), where students were assessed at the end of each school year with grade-specific assessments. The initial sampling of these students was such that 10 students from each grade were randomly selected and tested within each of the 60 schools. Once selected into the sample, these students were then followed for the duration of the panel until they reached the last grade surveyed, or they left the school for any reason. From the 3,000 unique students who were recruited as part of our study within the 2014-16 period, we end up with learning outcomes for 2833 of them. Since these schools did not

---

[7] For more information on where schools were located, and how these districts compare to the rest of the country, please see the Appendix ("c. More contextual details").

[8] We also have data for 2013, although due to changes in the assessments, and the lack of item-level data, we decide to only use it for robustness checks. In particular, in 2013 the difficulty of the grade 2 and 3 tests was deemed to not provide enough discriminating power. Therefore, the difficulty was adjusted for the following year, and then kept constant for the remaining three years.

receive any exclusive or targeted intervention that other school in different parts of the country did not also receive, the attrition in the sample does not threaten the representativeness of the dataset, as schools outside of the panel would have also been expected to display similar patterns of attrition. For most of the students in this panel, we also have information on household characteristics, and non-financial educational inputs at the household level – although these covariates were only collected in 2015 and 2016.

The tests measuring learning outcomes were designed and administered by Twaweza, who was simultaneously also in charge of the broader Uwezo initiative across East Africa which aims to document at-scale learning levels in foundational numeracy and literacy for children under 17. The tests were low-stakes exams, used purely for research purposes. Every year of the study, the students took a grade-specific test in math, English and Kiswahili. The test provided item-level data by subject (e.g., Kiswahili) and by sub-topic (e.g., reading words in Kiswahili) The assessments were similar across years, which was partly done by developing test booklets which kept the same items "in spirit" across years, but whose digits or words were modified each subsequent year. Appendix E shows examples of math and Kiswahili questions for all four years.

For our primary outcomes, we use the continuous test scores obtained from these assessments. We show two different scoring approaches, one scoring the tests as a raw percentage of the total number of items per subject, grade and year, and another using item-response (IRT) for each test at the level of the subject, grade, and year (e.g., English for grade 1 in 2014). We use the IRT scores as our main test scores because IRT scoring can place weights differentially by item to maximize discriminating power, but for the most part, none of our results are sensitive to this choice. We standardize these scores within subject and year for all three years.

As an additional robustness check, we also attempt to create a second set of outcome scores by incorporating the 2013 scores so that we are able to examine a longer pre-treatment trend. However, these scores come from assessments that were different between 2013 and the rest of the years, harming the comparability of these scores with those from other three years. Furthermore, we do not have access to item-level data for this year. So, to incorporate this additional baseline year, we use the actual test booklets to manually flag questions within the 2014-2016 test booklets that most resembled those asked in 2013 for all subjects and grades, with the goal of creating a "pseudo-2013" test out of the 2014-16 assessments. We then created a percentage score as an outcome for each grade, subject, and year, considering only those items that made the 2014-2016 most resemble the 2013 assessments. Finally, this outcome is also standardized by subject and grade against the pooled sample from all years. As an additional robustness check, we repeat this exercise using the 2014 booklets to find equivalent questions in 2015 and 2016 so that we also have access to a "pseudo-2014" measure. In a sense, this approach not only serves as a robustness

check by adding a baseline year (in the case of the "pseudo-2013"), but by ensuring the comparability of assessments across years by manually picking items that most resemble each other across time.

Leveraging the item-level data, we also create two other sets of outcomes of interest. First, we identify which specific sub-skills each student is mastering each year. In particular, we follow the approach of international assessments like Uwezo and determine that if a student can answer over half of all questions for a given sub-skill correctly, they are flagged as having mastered that sub-skill. Second, we leverage the outcomes for these sub-skills to label each student as having achieved specific "grade 1-" or "grade 2-proficiency" in each of the three subjects. We define "minimum grade-level proficiency" based on the curricular expectation pre-reform, and as such, these are mastering addition by grade 1, and multiplication by grade 2. For Kiswahili and English, these consist of reading sentences by grade 1, and reading paragraphs by grade 2. Both of these measures allow us to speak to policy effects on more concrete units of policy-relevance like grade-level proficiency and mastery of key numeracy and literacy skills.

In terms of school and teacher data, we have some information on school facilities, management practices, and school income and expenditures. Appendix Figure 1 describes these schools in our sample as of 2013. Enumerators also surveyed all teachers (about 1,500) who taught the students in our focal grades (grades 1, 2, 3) and focal subjects (math, English and Kiswahili), and collected data on individual teacher characteristics such as education and experience, as well as effort, teacher satisfaction, and teaching practices (e.g., whether teachers tried "tracking" within their classrooms).

Other achievement data

The main learning data from Mbiti et al. (2019) and Mbiti et al. (2021) has two key strengths in that (1) it has item-level information, which allows us to decompose treatment effects by sub-skills driving the changes, and (2) it samples the same schools and children across time, reducing the extent to which differences across time are simply due to random sampling variation. We complement this main learning data with two additional measures of student achievement that do not have these advantages but that do allow us to examine achievement trends over a longer period of time. First, we use Uwezo learning data from 2010-2017 (excluding 2016, as Uwezo data was not collected this year). Uwezo is a large-scale national, citizen-led data collection effort led by civil society organization Twaweza as a tool to benchmark learning outcomes in East Africa, including Tanzania through a low-stakes assessment that aims to be as representative of the whole country as possible. These data sets are publicly available, and cover children of roughly ages 5-17. Like our main outcomes, Uwezo tests cover English, Kiswahili, and math, and in fact, the test booklets administered for the Mbiti et al. studies (2019, 2021) are modelled after the Uwezo tests.

Uwezo data does not have item-level outcomes, but it rather places children at a given "level" for each sub-skill within each subject (e.g., student j is at the addition level in math, at the letter level in English, and at the syllable level in Kiswahili). We transform these outcomes into numeric scores that are comparable for all years across the 2010-2017 Uwezo panel, and which allow us to compare these outcomes with the outcomes from the main panel of learning outcomes[9]. This secondary data set allows us to increase the number of observations in the estimations, and to explicitly test for pre-trends – which the single pre-period in the main learning data does not allow for. Having said this, we give preference to the data from Mbiti et al. (2019) and Mbiti et al. (2021) because, again, Uwezo does not provide item-level data which allows for a more consistent grading of the outcome, is not necessarily sampled in a consistent manner across years, is not statistically guaranteed to be nationally representative, and consists of repeated cross-sections of data collection, increasing the risk of random noise affecting cross-year comparisons[10].

The second additional source of achievement data consists of test scores for national examinations in grade 4 ("Standard Fourth National Assessment" or SFNA) and in grade 7("Primary School Leaving Examination" or PSLE). These are publicly available at the individual-level at https://www.necta.go.tz, and contain information about the universe of students in Tanzania, allowing us to understand what happened to school enrollment by the time the students in our main panel reached grade 4. For the purposes of this analysis, we use scraped data on both tests from 2015-2018. The main goal of the grade 4 test scores is to understand the long-term effects that the policy reform had on grade 4 exam passing rates. The grade 7 students serve as a control group for the same period, given that even the oldest cohort to be affected by the policy would not have been in grade 7 until 2020, outside our period of study. The scores for both assessments are reported separately for each subject, and the outcomes are given in letter grades, where a student needs to score a C or above to pass the examination. Using these letter grades, we create binary variables flagging whether a child passed that subject or not. Given the anonymization of our main panel, we cannot link our initial learning panel with this administrative data base at the student level, but we still analyze these data at the level of student.

---

[9] We realize that, qualitatively, the discrete changes from certain sub-skills to higher skills may not represent the same underlying change in learning outcomes. In other words, it may be cognitively more challenging to go from the level "nothing" to "counting" than to go from "addition" to "subtraction". However, for the purposes of this analysis and additional robustness checks, we needed to create a unified continuous score using the Uwezo data, and as such, this was the most transparent approach.

[10] Furthermore, the potential categories in which children could have been placed within the Uwezo dataset changed from 2014-2015, only to then return to the previous system in 2017. More specifically, in 2015, some skills were broken into finer levels of disaggregation, such as addition becoming "single-digit addition" and "double-digit addition". This forces us to make assumptions in terms of how to reconcile this data cross years, and our results may be affected by these empirical decisions.

Empirical strategy

We exploit the variation in the timing of the policy introduction, and in the grades targeted by the curricular reform to estimate the impacts of the reform on learning outcomes through a difference-in-differences (DiD) framework. The intuition behind our identification strategy is that, absent the curriculum reform, the trend in performance for students in treated grades (1-2) would have remained similar to that of students in the untreated grade (3). Therefore, our preferred specification follows the structure of a classical two-period, two-group DiD strategy, like that found in Beatty and Shimshack (2011) and Carvalho and da Mota (2017). In this case, our first difference consists of the difference in learning levels, within each grade, before and after the reform. Our second difference is the difference between grades that were targeted by the reform (grades 1-2), and the grade that was not (grade 3), which is how we account for the "secular trend" in the specification. In other words, we look at grade 1-2 outcomes before and after the reform, and account for trends in how learning levels changed over the same time period for other grades untreated by the reforms using the grade 3 data. In Appendix Figure 2 we display the different groups that are part of the identification strategy.

In particular, we estimate the following model:

$$[1] \qquad \text{Outcome}_{ijgt} = \beta_0 + \beta_1(\text{Treatment*Post}_{gt}) + \lambda_g + \text{Post}_t + \varepsilon_{ijgt}$$

Where the "Outcome" refers to the learning or enrollment outcome for individual i, subject j, grade g in year t. We introduce grade-level fixed effects through λ_g, and Post_t is an indicator variable which equals 1 for 2015. The coefficient of interest is β_1, attached to the "Treatment*Post$_{gt}$" term. Specifically, this variable equals 1 only when a child is in grade 1 or 2 in 2015 and the year is 2015 (after the reform was implemented).

For our main specification, we focus only on 2014-2015 data. Note that given the panel nature of the data, students who were in grade 3 in 2016 were affected by the policy when they were in grade 2 in 2015. Therefore, we give preference to the data from 2014-15, as opposed to also including 2016. Including the 2016 data with the current specification would group a cohort that was actually treated into the comparison group and would "contaminate" our comparison group. More specifically, if the reform had positive effects on learning, this approach might yield underestimates of any potential increases in learning as a direct result of the policy.

Similarly, notice that in our current specification, the students in 2014 who constitute the comparison group for grade 1 are also those who are the treated group for grade 2 in 2015. Contrary to the case described in the paragraph before, we do not believe that this poses a threat to our identification strategy. This is because outcomes are observed at the end of each grade and therefore, the outcomes for this cohort when they were in grade 1 are observed in 2014 after completing grade 1 under the previous curriculum, and as such,

our specification correctly groups them into the comparison group for grade 1 – those who did not receive the 3R curriculum in grade 1. In the same manner, we observe their end-of-year outcomes for grade 2 in 2015 – after having completed grade 2 under the 3R curriculum, and hence, are properly classified as part of our treatment group by the current specification. A similar argument can be made for those students in grade 2 in 2014, who are part of the comparison group for grade 2 in 2014, and then become the group which contributes "post" information for grade 3 in 2015. In all, we do not believe that this issue poses a challenge to our internal validity, as this specification properly groups students into their corresponding treatment and comparison classifications within each year. Given the relatively small number of groups, we present both robust standard errors, and also p-values emerging from wild-bootstrapped clustered standard errors at the grade-level. In general, our results are not sensitive to the empirical decisions described here.

The difference-in-differences identification strategy requires that we justify whether the parallel trends assumption holds in this case. In other words, our main the assumption is that, absent the curriculum reform, students in a treated would have experienced similar trends in performance to untreated grades. In the case of the long-term analyses, this assumption would imply that, absent the reform, the number of test-takers and passing rates in grade 4 would have experienced the same trends as the number of test-takers and passing rates in grade 7 within the same year. Unfortunately, we cannot explicitly show parallel trends using our main data (that from Mbiti et al., 2019 and Mbiti et al., 2021), as we only have one year of consistent data from the period before the curriculum reform. We explore this issue by using 5 years of Uwezo data before the reform to visually check for differences across the different cohorts in this large-scale assessment. As mentioned before, Uwezo assessments in Tanzania are very similar to the instruments used to collect the data used in the current paper, as they were developed by the same organization, around the same time, and with the same aim of measuring foundational knowledge in math, English, and Kiswahili. We visually show in Figure 1 that students in lower grades (1-4) do seem to move in the same trajectory in all three subjects, for the 5 years of data available in the pre-period.

Qualitatively, we argue that the "spirit" of parallel trends might not be met in at least two cases. First, there could be another policy that heterogeneously affects one of the grades in the sample and hence confounds our estimates. As previously discussed, we are not aware of any other policy of the kind for these grades between 2014 and 2015. We also believe that the parallel trends assumption may not hold if there is a change in the composition of the incoming cohorts, which may introduce selection bias in the estimates of our treatment effects. For this specific case, we are aware that in 2016, the Tanzanian government introduced the Fee-Free Basic Education (FFBE) policy, which made primary education more accessible to students of more disadvantaged socioeconomic status in grade 1. Specifically, we observe in the data that the pupil-to-teacher (PTR) increased from 87 to

122 (40%) for grade 1 between 2015 and 2016. However, the PTR for grades 2 and 3 remains constant at 79 and 32 respectively over the same period. However, since we focus on the 2014-15 period, this reform does not affect our main estimates, nor do these cohorts reach grades 4 or 7 within the time window of our long-term analysis.

V.   Results

The reform improved foundational literacy and numeracy in grades 1-2

We find that the curriculum reform had a positive and statistically significant effect on math and Kiswahili learning outcomes one year after the reform. As the first row of Table 2 shows, students experienced an increase of 0.19 SD in an index outcome that combines the two main subjects targeted by the reform, and an increase of 0.20 SD in each of these subjects when estimated separately. As shown in the other two rows of Table 3, these results are directionally the same, and even of larger magnitude, if one uses secondary measures that attempt to increase the comparability of the assessments across years. Similarly, as Appendix Table 3 shows, these results are directionally the same when using outcomes from the Uwezo dataset, although the differences in how the outcomes are reported and the different time periods do change the magnitude of the treatment effects (in this case, decreasing them closer to 0.1 SD). Although the reform de-emphasized English instruction, we find no evidence of large reductions in English test scores in our main outcomes. Our point estimates are positive and the standard errors are such that we can rule out a negative effects smaller than -0.02 SD with 95% confidence. As we will explore further when we discuss the effects on sub-skills, we believe the improvement in English was due to spillover effects on basic skills transferrable from one language to the other.

Another way to understand these learning gains is to examine what happened to levels of minimum grade-level proficiency as a result of the policy. As shown in Figure 3, these results are not only meaningful in units of standard deviations but also in terms of reaching minimum proficiency levels. For instance, the policy reform increased the likelihood of a student reaching grade 1 math proficiency by 40%, and it more than doubled the likelihood of a student reaching grade 2 math proficiency. Similarly, it increased the probability of reaching grade 1 proficiency in Kiswahili by 29% and grade 2 by 71%. Even in English, the probability of reaching grade 2 proficiency increased by 7 percentage points over a base of 2%. The large magnitudes of these relative increases across all three subjects are partly due to the significant positive effects on learning, but also due to the low baseline levels of learning achieved by pupils in the sample, and in Tanzania more broadly.

In terms of attrition, —that is, pupils leaving our panel before we would expect them to given their grade— we estimate that the reform had a causal reduction in the attrition rate from the sample of 6 percentage points. Our data cannot track individual students across

the universe of Tanzanian schools, so we cannot definitively claim that this attrition is equivalent to school dropout. In other words, there are two potential interpretations for the reduction in attrition among treated students. The first hypothesis is that the reform made students more likely to remain in the schools where they were enrolled at the start of our data panel. Although plausible, the reform did not target specific schools, so we do not have an ex-ante reason to believe that the relative quality and desirability of schools changed because of the reform. The second interpretation, and the one we favor, poses that this decrease in attrition among treated students was indeed linked to a decrease in dropout. Particularly when this hypothesis is coupled with the results we present below on longer-term outcomes, it appears the policy not only led to improved learning, but also higher enrollment retention of students.

Skills across the range of complexity improved as a result of the reform

We would also like to understand whether the policy had heterogenous effects on the different literacy and numeracy sub-skills (e.g., "reading words in English") that were assessed. This is a valuable exercise as it can provide evidence on the mechanisms through which the reform operated. For example, did the reform only improve basic skills but weaken the more complex sub-skills? Or did help students master higher order concepts, but not at improving the more foundational skills? Since we have access to item-level data which we can aggregate up to the level of these sub-skills for each student, and we leverage the fact each grade was tested on very similar topics and skills across years and estimate the effect of the policy on each sub-skill. We use our main difference-in-differences specification to estimate the effect of the policy on the likelihood of mastering each of the sub-skills that students were tested on.

Figure 2 shows the estimates of the policy on specific sub-skills by subject. For math, it is not clear that the level of complexity of the sub-skills moderated how much the reform affected these tasks. In fact, sub-skills across the whole spectrum of complexity benefited from the policy (e.g., inequalities, addition, and multiplication). If at all, this figure shows that the reform indeed strengthened the most foundational sub-skills at a similar rate as the more complex tasks. Similarly, there were gains across the spectrum of complexity in Kiswahili. Together, these two findings suggest that the improvements in learning spurred by the reform did not come at the expense of sub-skills at either end of the spectrum: they did not "over-simplify" the instruction such that only the most foundational skills were improved, nor did it only benefit pupils already mastering a certain level of proficiency.

Interestingly, the two most basic English sub-skills that were assessed, meaning "recognizing letters" and "reading single words", also seem to have improved because of the policy. Even if the policy moved instruction away from English, if improvements in Kiswahili literacy were to have any spillover effects on other subjects, one would

hypothesize that they would be in the most basic literacy skills of another language which uses the same script, but not necessarily as much in higher order English skills, as we find here. As such, de-emphasizing English during these two first grades did not lead to overall losses in English learning. This was partly due to the low baseline levels shown in Table 3, as students were close to floor of the assessment at this point. However, we also have suggestive evidence that another potential mechanism for the lack of decreases in English test scores was by creating a common foundation in Kiswahili to build upon and, as such, the policy was also able to speed up the acquisition of more advanced skills in a different subject.

The policy led to higher enrollment, but lower passing rates four years after first implemented, likely due to compositional effects

Next, we explore whether the policy had persistent effects on educational outcomes. We leverage the universe of standardized national test scores from 2015-2018 for grades 4 and 7 grade to explore whether the curriculum reform also led to changes in educational attainment in the longer term, which we show in Table 4. In particular, the first cohort to be fully under the new curriculum is those students in grade 1 in 2015, and who were our treated group in grade 4 by 2018. Therefore, when using these data and our main model, our treatment group consists of repeated cross-sections of grade 4 students, with grade 7 students serving as the comparison group. The pre-period consists of the 2015-2017 period, and the post period 2018. Note that under this set up, those in grade 4 in 2017 were technically affected by the new curriculum when they were in grade 2. As such, the estimates that emerge from using the full sample here can be interpreted as underestimates of the true estimates. However, as a robustness check, we also display in the second row, the estimates resulting from the same specification but dropping 2017 for both grades 4 and 7 students, which would remove any potential (upwards) bias from the control group.

These results show two key patterns. First, the number of test-takers —a proxy for system-wide enrollment— increased by 16-17%. This is consistent with the decrease in attrition observed using the main panel data shown in Table 2. This result is also consistent with the hypothesis that learning and enrollment are linked to a certain extent, as either enrollment leads to higher learning (as Bau et al., 2021 might suggest), and/or higher learning leads to a higher likelihood to remain enrolled. Having said this, this increase in enrollment came with decreases in the passing rate of these national grade 4 examinations. In particular, using the baseline rates as a benchmark, the passing rate in math decreased 5-7%, the passing rate in English decreased 11-16%, and the passing rate in Kiswahili decreased 16-19%.

These long-term changes are suggestive of two facts. First, the increase in enrollment led to compositional changes in the universe of students reaching grade 4, particularly

towards the inclusion of lower-performing students who would have otherwise dropped out of school by grade 4. Second, these decreases in passing rates do not negate the short-term learning gains in learning observed. At worst, these decreases in performance are comparable in magnitude to the increase in student enrollment, which suggests that aggregate learning and educational attainment still increased –if one assumes that those who did not pass still learned something over these four years– relative to a counterfactual where the reform was not implemented and fewer children would have been enrolled in school.

Teacher training *may* have moderated gains

As described in Section III, rolling out nationwide curricular reforms in a large country like Tanzania is logistically challenging, and it is likely to yield heterogenous effects at the local- and individual-level due to variation in implementation across contexts. A key component of a curriculum reform of this scale is teacher training, as teachers must be aware and capable of implementing the expected instructional changes. In fact, weak teacher training was identified as one of the main reasons for the failure of a curricular reform aimed at improving early literacy outcomes in grade 1 in Costa Rica (Rodriguez-Segura, 2020). Therefore, we explore the extent to which teacher training may have moderated learning gains in this context. Teacher training was not randomly assigned at baseline, and its implementation varied across Tanzania depending largely on the regional entity in charge of imparting the training (Komba and Shukia, 2021). As such, we can only provide suggestive and correlational evidence for the issue of teacher training. Having said this, Table 6 shows the treatment effects of schools that had any teacher trained in the new 3R curriculum in 2014 - before the policy was actually implemented, and for those schools that did not. This table suggests that receiving teacher training was imprecisely correlated with larger treatment effects across the two subjects. Together, these results are suggestive that beyond informing teachers of a change in the allocation of time across subjects, training them on how to do it may be a key element to achieve larger learning gains through a reform of this type.

For the most part, more disadvantaged groups of children benefited more from the policy

We also study whether certain demographic characteristics are correlated with heterogeneous gains in learning, as shown in Table 7. We observe that female and rural students drove most of the treatment effects for learning. In other words, groups that are typically considered more disadvantaged in this context benefited the most from the policy in terms of learning in literacy and numeracy. We do not observe any heterogeneity by grade, as the difference between the two grades is not substantively or statistically significant. In terms of attrition, the patterns are similar except for the difference between

urban and rural students, as urban students drive most of the decrease in attrition. In all, while this sub-group analysis sheds light on some heterogeneous effects by demographic characteristics, it does not reveal that an exclusive sub-group benefited from the policy across the board. Instead, the gains seem to be, to some extent, distributed across different sub-groups, and if at all, they benefited disadvantaged groups more than their peers.

Other contemporaneous reforms do not appear to confound the effects

As described in Section III, this curriculum reform was only one part of a suite of reforms undertaken under the heading of Big Results Now - described in greater depth in the Appendix ("d. Description of other contemporary reforms"). None of these reforms targeted directly or differently our treated and comparison groups, and some of these reforms even happened after our period of analysis. However, we still empirically test whether we find some heterogeneity due to these reforms.

One of the other reforms that could be affecting our results is the Student Teacher Enrichment Programme (STEP - implemented in 2014). This policy trained teachers on how to identify struggling students and support them. The STEP training was rolled out in selected districts, and 4 out of 10 of our districts were in this group. We run our main specification only on districts that were not STEP districts, and display this in Table 7. When broken down by whether a district was part of the STEP program, the results similar and the difference is not statistically significant. Therefore, it does not appear that the implementation of the STEP program is confounding our main treatment estimates.

There were two other school-wide reforms for which we check whether we have heterogeneous effects: the distribution of School Improvement Kits (including the "Mwaongozo" leadership training for head teachers), and the school grants disbursed by the Tanzanian government. The former deals with the quality of school management, and the latter deals with a fairer distribution system of school funding. Table 7 again shows the heterogeneity results for both of these school characteristics. Although the differences between the groups do not rise to be statistically significant, the magnitudes of the differences are medium-sized. Much like Mbiti et al. (2019) and Mbiti et al. (2021), we believe that, if these differences are indeed suggestive of treatment effect heterogeneity, the presence of adequate school resources may have augmented the effectiveness of the curricular reform, but not necessarily confounded the treatment effects, as none of these policies were targeted at specific grades. It is also worth noting that the implementation of most of BRN components were delayed due to the lack of funding. For instance, the capitation grant reform was only launched in 2016, the last period of our study.

Finally, these schools in the core sample of the paper were explicitly chosen as control schools in the companion experimental evaluation, so by default, we ensure that they were not affected by the other interventions being rolled out by researchers. Other

reforms, such as the school ranking program was the first component launched and one of the few that was consistently implemented throughout our study period. However, this program focused on results in grade 7, which would be completely out of reach for even our oldest cohort.

VI.   Discussion

Our results suggest that the Tanzanian curricular reform of 2015 improved foundational literacy and numeracy for early grade students. The targeted restructuring of instruction within an overcrowded curriculum, coupled with low achievement levels at baseline, led to significant improvements in proficiency levels for early literacy and numeracy. These results are robust to the use various selections of items in the assessment, different definitions of the outcome variables, and do not seem to be fully driven by any of the other Big Results Now reforms. These findings provide empirical backing for the prior set forth by papers like Pritchett and Beatty (2015) or Muralidharan et al. (2019), which advocate for a realignment and simplification of curricula in LMIC to allow students to properly develop early literacy and numeracy. These results also describe a successful case study where such a reform was implemented and led at the national level by the government of a LMIC like Tanzania.

The strengthening of the foundational numeracy and literacy skills through this curricular reform in the earlier grades highlights the key role that curriculum design plays as a key input for educational systems. In particular, these results challenge policymakers to explore the sharpening of curricula in developing countries, and their specific targets during the earlier years of education. In a sense, the poor learning outcomes in developing countries need not be fully explained by irreversible school and student characteristics, but also by the pedagogy of how the material is taught, and what is expected of students. Failing to meet curricular standards could be both due to the student's low levels of learning, but also due to the stringent, overambitious, and unrealistic standards that they are subject to, both through fast-paced instruction and unrealistic assessments. Interventions such as Teach at the Right Level (for instance, see Banerjee et al, 2017) or the current study show that thoughtful curricular design and pacing can lead to promising gains in learning.

We also find that the curriculum reform led to increased school enrollment, at least until grade 4. The decrease in student dropout as a result of a curriculum reform which focuses on strengthening FLN is both a welcome and unsurprising effect. A curriculum that is more tailored to most students' needs and does not focus (as much) on high performing students —likely from a high socioeconomic background— is likely to have larger effects on students that were more likely to leave school prematurely, as shown in the current study. Yet, this decrease in school dropout does not imply that the educational system has done its part with these new entrants. We also observe decreases in the passing rate of the grade

4 national examination that matches very closely the increase observed in school enrollment. This fact suggests that even if the reform did lead to higher enrollment and learning gains in the short term for FLN, these gains were not enough for most of these new entrants to pass the grade 4 national examinations. These new students are likely to be from more disadvantaged backgrounds, and while the curriculum reform likely aligned classroom instruction closer to their achievement level, it may have not met all the educational needs of these children. Therefore, these results are indicative that while this type of curriculum reform may be beneficial and desirable, it may not be enough to ensure educational success in the long-run. Additional interventions, such as more individualized instruction or the revision of the curriculum of the higher grades as well, may be needed to help these children keep succeeding later in their educational path.

Similarly, the effectiveness of a new curriculum, as well-designed as it may be, is likely to be dampened if all of parts of said educational system are not aligned to work well with this change, that is, if "implementation" is poor. For instance, in the case of the 3R curriculum reform, training even a single teacher per school ahead of implementing the reform was correlated with larger gains in learning. This is suggestive that, unsurprisingly, the quality of implementation of a new curriculum can moderate the effects of curricular reform policies. The World Bank makes this point on their Report on Learning: "if a country adopts a new curriculum that increases emphasis on active learning and creative thinking, that alone will not change much. Teachers need to be trained so that they can use more active learning methods, and they need to care enough to make the change because teaching the new curriculum may be much more demanding than the old rote learning methods" (World Bank, 2017). Even in LMIC with weak state capacity, well-designed, and well-implemented, programs can greatly improve literacy outcomes in developing countries (for instance Kerwin and Thornton, 2019; or Eble et al., 2020) if part of the intervention design involves getting teachers to meet children at their level and gradually teach from there. We display a case in which a well-designed policy had the intended results in learning gains on average, but also where the results were likely magnified, at least correlationally, through better implementation at the local level.

Our study has several shortcomings and limits to what can be inferred from these results. First, our main results cover a very short time period. Therefore, we cannot ensure that the parallel trends assumption holds in the same data from which we draw our our main estimates. To address this, we use a different data set, Uwezo, and qualitative knowledge of the context to justify why parallel trends might hold. However, these options are only second best to a more comprehensive check for parallel trends in the same data set as the one we use for our treatment effects estimates. Second, the learning data collected at the beginning of each year was of poor quality, and as such, we cannot provide direct evidence on the heterogeneity of the effects by baseline performance. This is a valuable area for future

research to explore, as a potential worry with this type of curriculum reform is that it may affect high-performing students at the expense of low-performing students. Finally, we do not have strong metrics for the quality of implementation of the reform in each school, beyond information about teacher training on the new curriculum. Although we present some suggestive evidence that the quality of the implementation may moderate the effects of the policy, further research is needed on this issue, especially given other evidence (Komba and Shukia, 2021) highlighting that the policy was heterogeneously implemented across the country.

In all, our findings contribute to the literature on curricular reform in developing countries. More broadly, our results speak to the issue of adapting antiquated and "overambitious" curricula in developing countries to the current educational needs. Curricula affect all students within an educational system, and as such, well-designed and well-implemented curriculum reforms can be a valuable tool to boost educational outcomes at scale. The current study presents evidence of such a reform which was indeed successful, yet not perfect, at improving learning in a LMIC like Tanzania.

VII.   References

Atuhurra, J., Alinda, V. (2018). Basic education curriculum effectiveness in East Africa: A descriptive analysis of primary mathematics in Uganda using the "Surveys of Enacted Curriculum''. MPRA Paper No. 87583, University Library of Munich, Germany.

Atuhurra, J. and Kaffenberger, M. (2020). System (In)Coherence: Quantifying the Alignment of Primary Education Curriculum Standards, Examinations, and Instruction in Two East African Countries. RISE Working Paper Series. 20/057. https://doi.org/10.35489/BSG-RISE-WP_2020/057

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives, 31.(4), 73-102. https://doi.org/10.1257/jep.31.4.73

Banerjee, A. V., & Duflo, E. (2012). *Poor economics: A radical rethinking of the way to fight global poverty*. (Paperback first published). PublicAffairs.

Bau, N., Das, J., & Yi Chang, A. (2021). New evidence on learning trajectories in a low-income setting. International Journal of Educational Development, 84, 102430. https://doi.org/10.1016/j.ijedudev.2021.102430

Beatty, T. K. M., & Shimshack, J. P. (2011). School buses, diesel emissions, and respiratory health. Journal of Health Economics, 30(5), 987–999. https://doi.org/10.1016/j.jhealeco.2011.05.017

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. Journal of Public Economics, 168, 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007

Brunette, T., Piper, B., Jordan, R., King, S., & Nabacwa, R. (2019). The impact of mother tongue reading instruction in twelve Ugandan languages and the role of language complexity, socioeconomic factors, and program implementation. Comparative Education Review, 63. (4), 591-612. https://doi.org/10.1086/705426

Bryan, G., Chowdhury, S., Mobarak, A. M. Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh. (2014). Econometrica, 82(5), 1671–1748. https://doi.org/10.3982/ECTA10489

Crépon, B., Ferracci, M., & Fougère, D. (2012). Training the Unemployed in France: How Does it Affect Unemployment Duration and Recurrence? Annals of Economics and Statistics, 107/108, 175. https://doi.org/10.2307/23646576

Chicoine, L. (2019). Schooling with learning: The effect of free primary education and mother tongue instruction reforms in Ethiopia. Economics of Education Review. 69, 94-107. https://doi.org/10.1016/j.econedurev.2019.01.002

Chiplunkar, G., Dhar, Nagesh, R. (2020). Too little, too late: improving post-primary learnin outcomes in India. wORKING PAPER. https://riseprogramme.org/sites/default/files/inline-files/Dhar.pdf

Chisholm, L., & Leyendecker, R. (2008). Curriculum reform in post-1990s sub-Saharan Africa. International Journal of Educational Development, 28. 2, 195-205. https://doi.org/10.1016/j.ijedudev.2007.04.003

Datzberger, S. (2018). Why education is not helping the poor. Findings from Uganda. World Development. 110, 124-139. https://doi.org/10.1016/j.worlddev.2018.05.022

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. Science, 357 (6346), 47-55. https://doi.org/10.1126/science.aal4724

Ejdemyr, S., Kramon, E., & Robinson, A. L. (2018). Segregation, ethnic favoritism, and the strategic targeting of local public goods. Comparative Political Studies. 51(9), 1111–1143. https://doi.org/10.1177/0010414017730079

Erling, E., Adinolfi, L., & Hultgren, A. (2017). Multilingual classrooms: Opportunities and challenges for English medium instruction in low and middle income contexts.

Evans, D., & Yuan, F. (2019). Equivalent years of schooling: A metric to communicate learning gains in concrete terms. World Bank Policy Research Working Paper, 8752. https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-8752

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. American Economic Journal: Applied Economics. 1(1), 112-135. https://doi.org/10.1257/app.1.1.112

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries. In Handbook of the Economics of Education. (Vol. 5, pp. 653-743). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Government of Tanzania. "Big Results Now! Annual report 2013/14," Presidential Delivery Bureau 2015.

Jones, S., Schipper, Y., Ruto, S., & Rajani, R. (2014). Can Your Child Read and Count? Measuring Learning Outcomes in East Africa. Journal of African Economies, 23(5), 643–672. https://doi.org/10.1093/jae/eju009

Kaffenberger, M. (2019). PISE-D Reveals Exceptionally Low Learning. RISE Programme. Blog post. https://riseprogramme.org/blog/PISA-D_low_learning

Kaffenberger, M., Sobol, D., Spindelman, D. (2021). The role of low learning in driving dropout: a longitudinal mixed methods study in four countries. RISE Working Paper Series. 21/070. https://doi.org/10.35489/BSG-RISE-WP_2021/070

Kaffenberger, M., & Pritchett, L. (2021). A structured model of the dynamics of student learning in developing countries, with applications to policy. International Journal of Educational Development, 82, 102371. https://doi.org/10.1016/j.ijedudev.2021.102371

Kerwin, J. T., & Thornton, R. L. (2020). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. The Review of Economics and Statistics, 1–45. https://doi.org/10.1162/rest_a_00911

Komba, A. and Shukia, R. 2021. Accountability Relationships in 3Rs Curriculum Reform Implementation: Implication for Pupils' Acquisition of Literacy and Numeracy Skills in Tanzania's Primary Schools. RISE Working Paper Series. 21/065. https://doi.org/10.35489/BSG-RISE-WP_2021/065

Laitin, D. D., Ramachandran, R., & Walter, S. L. (2019). The legacy of colonial language policies and their impact on student learning: evidence from an experimental program in Cameroon. Economic Development and Cultural Change, 68(1), 239-272. https://doi.org/10.1086/700617

Lima, R. C. de A., & Silveira Neto, R. da M. (2018). Secession of municipalities and economies of scale: Evidence from Brazil. Journal of Regional Science, 58(1), 159–180. https://doi.org/10.1111/jors.12348

Malisa, M., & Missedja, T. Q. (2019). Schooled for servitude: the education of African children in British colonies, 1910-1990. Genealogy, 3(3), 40. https://doi.org/10.3390/genealogy3030040

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: experimental evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627-1673. https://doi.org/10.1093/qje/qjz010

Mbiti, I., Romero, M., Schipper, Y. (2021). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. NBER Working Paper No. 25903

Ministry of Education, Science and Technology,. (2016). Curriculum for basic education Standard I and II. ISBN. 978 - 9976 - 61- 436 - 7.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in India. American Economic Review, 109(4), 1426-1460. https://doi.org/10.1257/aer.20171112

Muralidharan, K., Sundararaman, V. (2011). Teacher performance pay: experimental evidence from India. Journal of Political Economy, 119(1), 39-77. https://doi.org/10.1086/659655

Mwiria, K. (1991). Education for subordination: African education in colonial Kenya. History of Education, 20(3), 261-273. https://doi.org/10.1080/0046760910200306

Piper, B., Zuilkowski, S. S., Kwayumba, D., & Oyanga, A. (2018). Examining the secondary effects of mother-tongue literacy instruction in Kenya: Impacts on student learning in English, Kiswahili, and mathematics. International Journal of Educational Development, 59, 110-127. https://doi.org/10.1016/j.ijedudev.2017.10.002

Pritchett, L., & Beatty, A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. International Journal of Educational Development, 40, 276–288. https://doi.org/10.1016/j.ijedudev.2014.11.013

Pritchett, L., Woolcock, M., & Andrews, M. Capability Traps? The mechanisms of persistent implementation failure. Center for Global Development. CDG Working Paper. (232)

Pritchett, L. (2013). The rebirth of education: Schooling ain't learning. Center for Global Development.

Sa, E. (2007). Language Policy for Education and Development in Tanzania. Working paper. https://www.swarthmore.edu/sites/default/files/assets/documents/linguistics/2007_sa_eleuthera.pdf

Slavin, Stuart, and Marcel F D'Eon. "Overcrowded curriculum is an impediment to change (Part A)." *Canadian medical education journal* vol. 12,4 1-6. 14 Sep. 2021, doi:10.36834/cmej.73532

SDI – Service Delivery Indicators. (2010-2014). Education data [Data file]. https://www.sdindicators.org/aboutus#where-are-we-now-

Seid, Y. (2019). The impact of learning first in mother tongue: Evidence from a natural experiment in Ethiopia. Applied Economics, 51(6), 577–593. https://doi.org/10.1080/00036846.2018.1497852

Twaweza. (2015). Uwezo data-Household data. [Data file]. https://www.twaweza.org/go/uwezodatasets

Uwezo. (2017). Are Our Children Learning? Uwezo Tanzania Sixth Learning Assessment Report, Dar es Salaam: Twaweza East Africa.

USAID. (2015). Testing and reading, writing and arithmetic (3Rs) reform in Tanzania. https://wwww.usaid.gov/sites/default/files/documents/1865/Bruns.pdf

Williams, T. P. (2017). The political economy of primary education: lessons from Rwanda. World Development, 96, 550-561. https://doi.org/10.1016/j.worlddev.2017.03.037

World Bank. (2017). World Development Report 2018: Learning to Realize Education's Promise. The World Bank. https://doi.org/10.1596/978-1-4648-1096-1

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 4, World Bank, Washington D.C. 2016.

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 5, World Bank, Washington D.C. 2016.

World Bank, "Tanzania Education Program for Results," Implementation Status and Results Report P147486 No. 2, World Bank, Washington D.C. 2015.

World Bank. (2019). "Tanzania Mainland Poverty Assessment: Tanzania's path to poverty reduction and pro-poor growth. Part I". http://documents.worldbank.org/curated/en/254411585030305188/pdf/Part-1-Path-to-Poverty-Reduction-and-Pro-Poor-Growth.pdf

VIII. Tables and Figures

Table 1: Estimated time allocation in hours per week across subjects and grades before and after the reform

	Grade	Math	Kiswahili	English
2014 (pre-reform)	1	3.3	3.1	2.9
	2	3.0	3.0	3.0
	3	4.1	4.4	4.0
2015 (post-reform)	1	3.6	5.0	1.9
	2	3.6	5.0	1.9
	3	4.1	4.2	4.2
Percent change from 2014 to 2015	1	9.1%	61.2%	-35.5%
	2	20.0%	66.7%	-36.7%
	3	0.0%	-4.5%	5.0%

Notes: figures derived from data on class observations by external enumerators.

Table 2: regression estimates of the causal effect of the curriculum reform on learning and enrollment

	Aggregate, all three	Aggregate, Math+ Kiswahili	Math	English	Kiswahili	Attrition
Main estimates (IRT)	0.16**	0.19**	0.20**	0.14	0.20*	-0.06**
	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)	(0.02)
	[0.14]	[0.19]	[0.17]	[0.33]	[0.17]	[0.12]
	3275	3275	3132	3132	3132	3023
Pseudo-2014 test	0.16**	0.19**	0.18**	0.12	0.23*	-
	(0.06)	(0.06)	(0.05)	(0.09)	(0.1)	
	[0.13]	[0.18]	[0.04]	[0.4]	[0.18]	
	3275	3275	3132	3132	3132	
Pseudo-2013 test	0.20**	0.25**	0.29**	0.13	0.25**	-
	(0.06)	(0.07)	(0.08)	(0.09)	(0.08)	
	[0.04]	[0.09]	[0.15]	[0.4]	[0.41]	
	3275	3275	3132	3132	3132	

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Table 3: regression estimates of the causal effect of the curriculum reform on achieving minimum proficiency levels of grades 1 and 2

	Math		English		Kiswahili	
	Control group	Estimates	Control group	Estimates	Control group	Estimates
Achieving G1 minimum proficiency	0.47	0.19***	0.03	0.04	0.41	0.12**
	(0.50)	(0.04)	(0.18)	(0.02)	(0.49)	(0.04)
		[0.18]		[0.17]		[0.14]
	N=1066	N=3132	N=1087	N=3132	N=1087	N=3132
Achieving G2 minimum proficiency	0.10	0.17***	0.02	0.07***	0.21	0.15***
	(0.30)	(0.03)	(0.13)	(0.01)	(0.41)	(0.03)
		[0.12]		[0.12]		[0.15]
	N=1066	N=3132	N=1087	N=3132	N=1087	N=3132

Notes: coefficients from linear probability model. Robust standard errors of coefficients, and standard deviations of control group in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Table 4: regression estimates of the causal effect of the curriculum reform on learning and enrollment

	Passing rate in math pre-reform	Math	Passing rate in English pre-reform	English	Passing rate in Kiswahili pre-reform	Kiswahili	Number of takers pre-	Students in 4th grade
Main estimates	0.45	-0.03***	0.38	-0.06***	0.8	-0.13***	1,100,868	173,014***
		(0.00)		(0.00)		(0.00)		(55.7)
		[0.21]		[0.16]		[0.27]		[0.16]
		N=8,062,450		N=8,062,450		N=8,062,450		N=8,062,450
Main estimates, no 2017	0.41	-0.02***	0.36	-0.04***	0.78	-0.15***	1,048,686	176,784***
		(0.00)		(0.00)		(0.00)		(8.55)
		[0.21]		[0.22]		[0.16]		[0.27]
		N=5,953,025		N=5,953,025		N=5,953,025		N=5,953,025

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Table 5: comparison of point estimates of schools that received
some teacher training in 2014, and those that did not

	Aggregate (all three)	Aggregate (Math+ Kiswahili)	Attrition
No teachers trained in 2014	0.13*	0.14*	-0.06**
	(0.06)	(0.07)	(0.02)
	[0.00]	[0.00]	[0.00]
	3076	3076	3076
At least one teacher trained in 2014	0.31*	0.37**	-0.06
	(0.13)	(0.14)	(0.04)
	[0.25]	[0.26]	[0.00]
	771	771	771
Difference no teachers trained-at least one teacher trained	-0.18 (p=0.20)	-0.23 (p=0.14)	-0.01 (p=0.86)

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses.
Wild-bootstrapped p-values in squared parentheses. Significance levels, based on
robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Table 6: Heterogeneity of results by different baseline and demographic characteristics

	Female	Male	Female-Male	Rural	Urban	Rural-Urban	G1	G2	G1-G2
Aggregate learning index (Math+Kiswahili)	0.23**	0.15*	-0.08	0.25***	-0.14	0.39***	0.17**	0.20***	-0.03
	(0.09)	(0.09)	p=0.55	(0.07)	(0.11)	p=0.00	(0.07)	(0.07)	p=0.64
	[0.15]	[0.15]		[0.14]	[0.64]		[0.16]	[0.16]	
	1641	1634		2729	546		2174	2197	
Attrition	-0.10***	-0.02	-0.08**	-0.04**	-0.16***	0.12**	-0.07***	-0.05**	-0.02
	(0.02)	(0.03)	p=0.03	(0.02)	(0.05)	p=0.04	(0.02)	(0.02)	p=0.16
	[0.05]	[0.37]		[0.13]	[0.16]		[0.16]	[0.16]	
	1641	1634		2729	546		2174	2197	

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels,
based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Table 7: Heterogeneity of results by whether schools were affected by other contemporaneous reforms or policies

	STEP	No STEP	STEP-No STEP	School Improvement Kit	No school Improvement Kit	Kit-No Kit	Below median financial support	Above median financial support	Below-Above
	0.20*	0.18*	-0.02	0.20**	0.09	0.11	0.12	0.26**	-0.14
	(0.09)	(0.09)	p=0.89	(0.07)	(0.17)	p=0.54	(0.09)	(0.09)	p=0.28
Aggregate learning index (Math+Kiswahili)	[0.38]	[0.12]		[0.18]	[0.13]		[0.08]	[0.18]	
	1313	1962		2796	479		1714	1561	
	-0.10**	-0.04	-0.07*	-0.06**	-0.07	0.01	-0.05	-0.08**	0.04
	(0.03)	(0.02)	p=0.07	(0.02)	(0.05)	p=0.86	(0.02)	(0.03)	p=0.32
	[0.08]	[0.27]		[0.14]	[0.13]		[0.12]	[0.13]	
Attrition	1313	1962		2796	479		1714	1561	

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

.

Figure 1: visual display of parallel trends using Uwezo data set
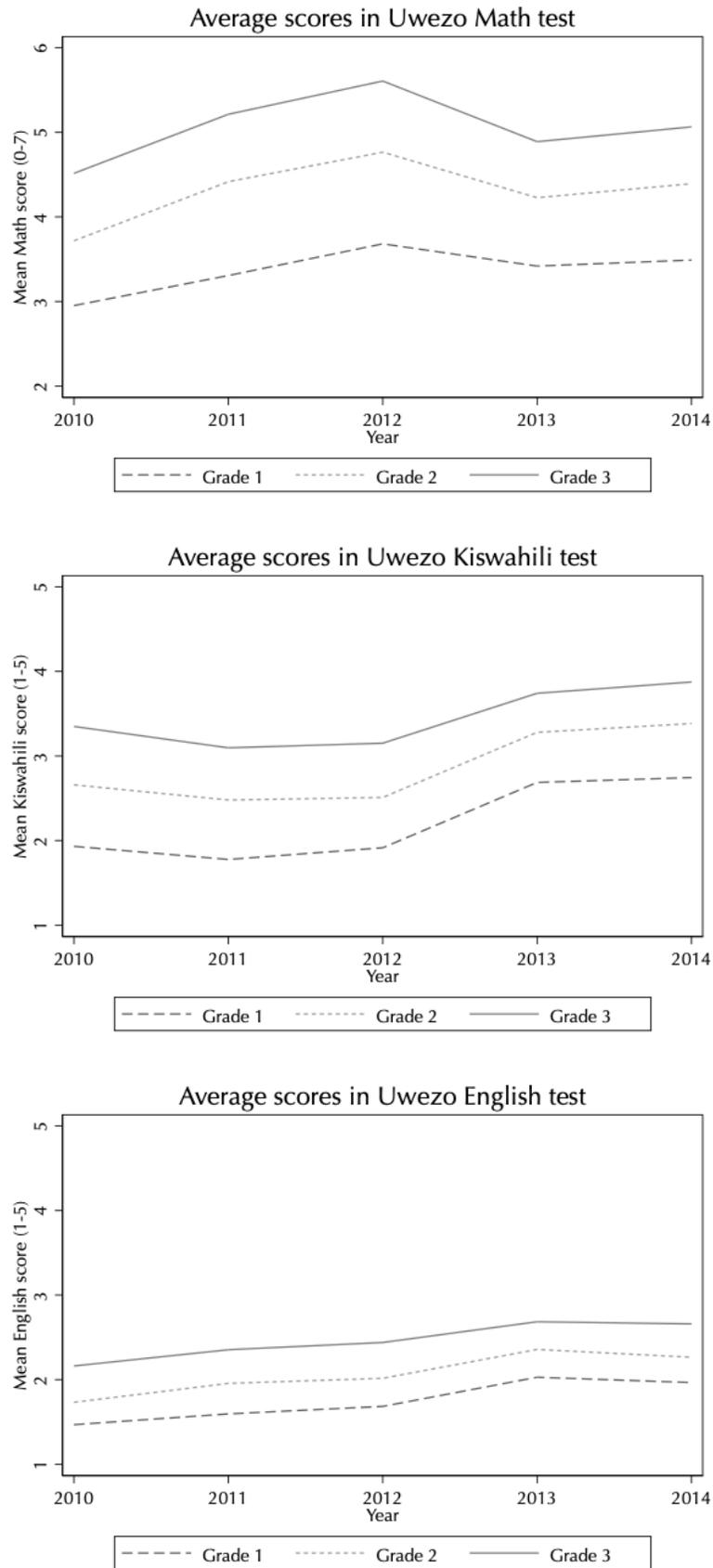
Figure 2: Comparison between treated and control cohorts in the probability of mastering different sub-skills

Math sub-skills

English sub-skills

Kiswahili sub-skills
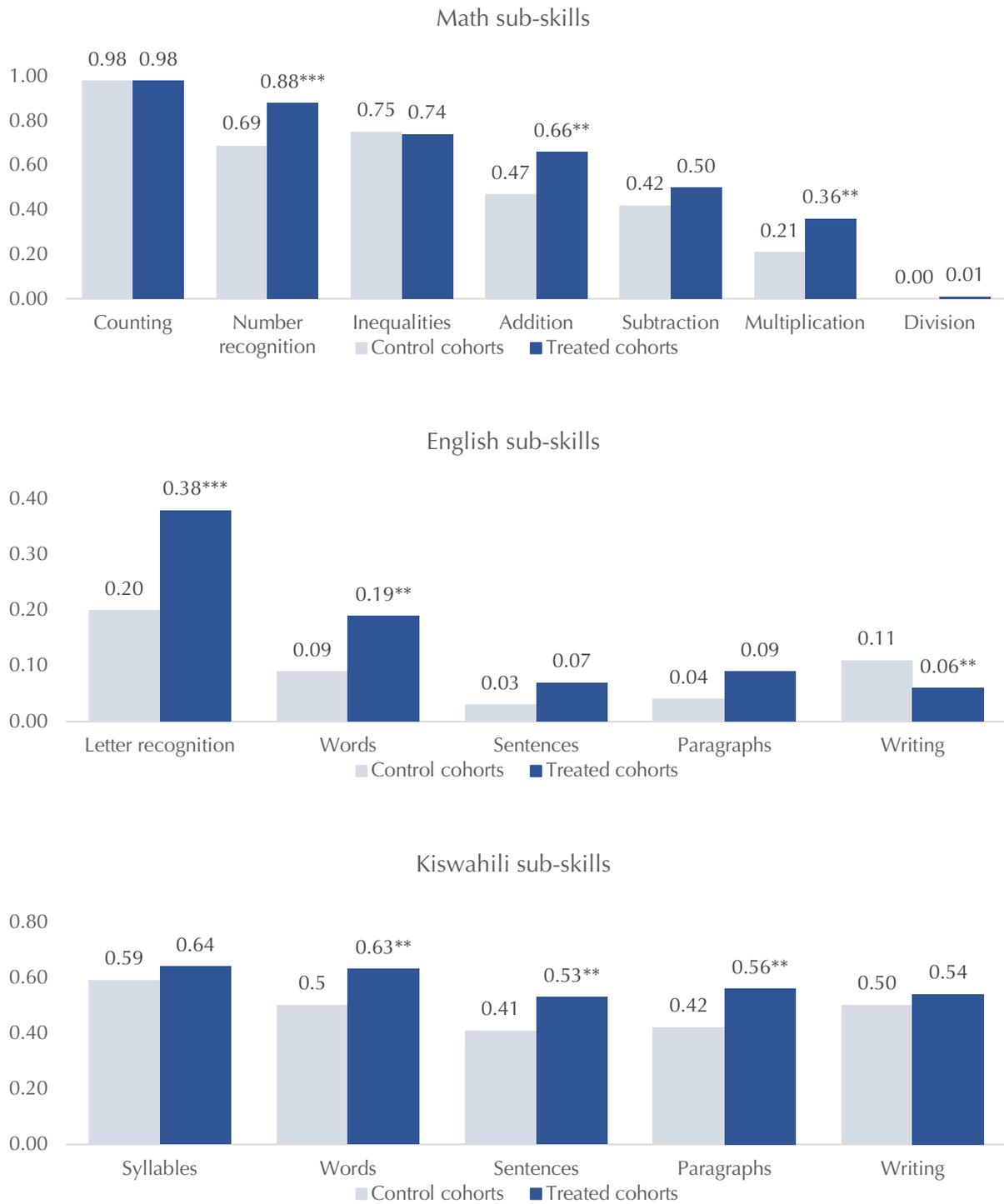
IX.   Appendix

a. Additional figures

Appendix Figure 1: summary statistics for school characteristics

Variable	Mean	SD
How many times did the committee meet last year	5.0	2.8
How many times did the Ministry of Education visit the school last year	2.4	2.8
How many teachers are there in this school	15.3	12.2
Is there an unfulfilled request for teachers	0.8	0.4
How many teachers are absent from the school at the moment	3.1	5.0
Is this an urban school	0.2	0.4
How many classrooms does this school have	9.0	3.8
Are there any unused classrooms or vacant rooms in the school	0.2	0.4
Is overcrowding a problem at this school	0.9	0.3
Does this school provide breakfast	0.1	0.3
Does this school provide Lunch	0.2	0.4
Observations	60	

Notes: averages calculated at the school level for 2013

Appendix Figure 2: structure of group cells for difference-in-differences

			Year	
			Pre period (2014)	Post period (2015)
Grade	1 2	Treatment grades (T)	T-pre	T-post
	3	Comparison grade (C)	C-pre	C-post

Notes: We highlight in green the group of students for which the interaction term between Treatment*Post would equal 1.

Appendix Figure 3: regression estimates of the causal effect of the curriculum reform on learning using Uwezo data (2010-2017)

	Math	English	Kiswahili
Main estimates	0.11***	-0.04***	0.01
	(0.01)	(0.01)	(0.01)
	[0.08]	[0.06]	[0.92]
	391,813	389,974	396,666
Estimates excluding grades 3 and 4	0.17***	-0.06***	0.10***
	(0.01)	(0.01)	(0.01)
	[0.00]	[0.00]	[0.44]
	275,193	273,965	278,760

Notes: these specifications exclude 2016, as Uwezo data was not collected for Tanzania in 2016. By 2017, grades 3 and 4 had been treated, so excluding these groups from the specification is the cleanest approach to avoid contamination of the comparison group. coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

Appendix Figure 4: Lee bounds for long-term effects on SFNA passing rates in 2018

	Main estimates		Main estimates, no 2017	
	Lower estimate	Upper estimate	Lower estimate	Upper estimate
Kiswahili	-0.13***	-0.12***	-0.15***	-0.14***
	(0.00)	(0.00)	(0.00)	(0.00)
	[0.16]	[0.16]	[0.00]	[0.16]
	N=7,494,870	N=7,960,261	N=5,541,256	N=7,494,870
English	-0.06***	-0.06***	-0.04***	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)
	[0.16]	[0.16]	[0.21]	[0.05]
	N=7,491,304	N=7,961,820	N=5,535,941	N=5,872,576
Math	-0.04***	-0.03***	-0.03***	-0.02***
	(0.00)	(0.00)	(0.00)	(0.00)
	[0.16]	[0.16]	[0.27]	[0.16]
	N=7,493,808	N=7,962,156	N=5,540,742	N=5,872,439

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Wild-bootstrapped p-values in squared parentheses. Significance levels, based on robust standard errors * p<0.10, ** p<0.05, ***p<0.01

b. What changed within classrooms and schools?

Beyond the change in curriculum, we also explore whether the reform was correlated with other behavioral responses at the school and classroom level. To do so, we group several variables which are available for all years either from head teacher, principal, and teacher surveys, and explore the differences before and after the policy reform. In the aggregation of these variables, we picked all the variables for which we have consistent and reliable data across pre- and post- years. We classified all these variables into these four, admittedly arbitrary, categories. Then, each variable was indexed from 0-1, where 1 was the "most positive" outcome of the variable. Each category consists of the geometric mean of the indexed version of each variable within it. Note that results hold whether we subset only to control schools or all 350 schools. The current results as displayed are just for control schools. The variables within each category were:

Effort and planning:
- Hours spent teaching in a week
- Hours spent planning lessons
- Hours spent managing and supporting teachers
- Personally taught remedial classes?

Instructional methods:
- Tried a new method of strategically assigning students in groups (tracking)
- Tried a new method of strategically assigning teachers to grades
- Tried a new method of more strictly enforcing student attendance
- Tried a new method of having more teacher supports (volunteers or trainee teachers)

Inputs and monitoring:
- Number of parent-teacher meetings this year
- Number of times Ministry visited the school this year
- (Inverse of) Whether the school holds any classes outside
- Amount of inputs compared to previous years

Teacher training:
- Amount of training compared to previous years
- Training of members of school committee

It is worth noting that current data limitations for these specific surveys only allow us to explore pre- and post- changes for this specific analysis of classroom and school changes, without a clear causal framework. Appendix Figure 5 shows the changes in the post period for the aggregated categories. The only two statistically significant categories are the one that reflects whether teachers are trying out new instruction methods, and the one showing the amount of training that teachers and head teachers got. This could be consistent with a story that beyond the implementation of the curricular change, teachers were better trained and hence did not need to try new pedagogical methods.

Appendix Figure 5: Regression results of changes in school and teacher characteristics after 2014

	Effort and planning	Instructional methods	Inputs and monitoring	Teacher and training
After 2014	0.045	-0.068	0.057	0.095**
	(0.03)	(0.03)	(0.09)	(0.04)
Observations	164	217	126	127

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Significance levels, based on robust standard errors * $p<0.10$, ** $p<0.05$, ***$p<0.01$

Contrary to the data limitations using the previous outcomes, we can analyze the self-reported teacher satisfaction on different issues using a difference-in-differences framework. Specifically, the treated group are teachers of grades 1-2, and the post- period consists of years 2015. The outcome variable is a discrete variable from 0 to 4, where 4 is the highest level of satisfaction reported on each issue. Appendix Figure 6 displays the coefficient of interest of each outcome. Although none of these coefficients emerges as statistically significant, the two largest coefficients by far are the satisfaction with self-perceived prospects for promotion, and school support. Both of these agree with the story that teachers realize their performance is improving, and that there are external factors beyond teachers that are facilitating this change.

Appendix Figure 6: Regression results using DiD estimator of
curriculum change on satisfaction ratings by issue

Satisfaction with…	DiD estimate
Current job	0.02
	(0.09)
Government support	-0.01
	(0.11)
Job security	-0.03
	(0.10)
Parental support	-0.02
	(0.10)
Promotion prospects	0.11
	(0.11)
Salary	0.03
	(0.10)
School support	0.18
	(0.11)
Observations	1371

Notes: coefficients standardized as z-scores. Robust standard errors in parentheses. Significance levels, based on robust standard errors * $p<0.10$, ** $p<0.05$, ***$p<0.01$

c. More contextual details

The map below shows in red the districts where the 60 schools in the sample were drawn from. Specifically, the districts are Geita, Kahama, Karagwe, Kinondoni, Kondoa, Korogwe, Lushoto, Mbinga, Mbozi, Sumbawanga, Kigoma, Kigoma, and Korogwe. According to World Bank poverty estimates at the district level (World Bank, 2019), the poverty rate at the district-level for the selected units is 27.3% with a standard deviation of 9.6%, comparable with the national poverty rate averaged at the district-level of 29.5% with a standard deviation of 13.9%. The average 2013 rank on the Primary School Leaving Examination, a standardized test taken in grade 7, for the districts in the sample is 71.4 out of 151, with a standard deviation of 47.6, a range that covers generally the national median and mean, and also represents districts in both ends of the achievement distribution.
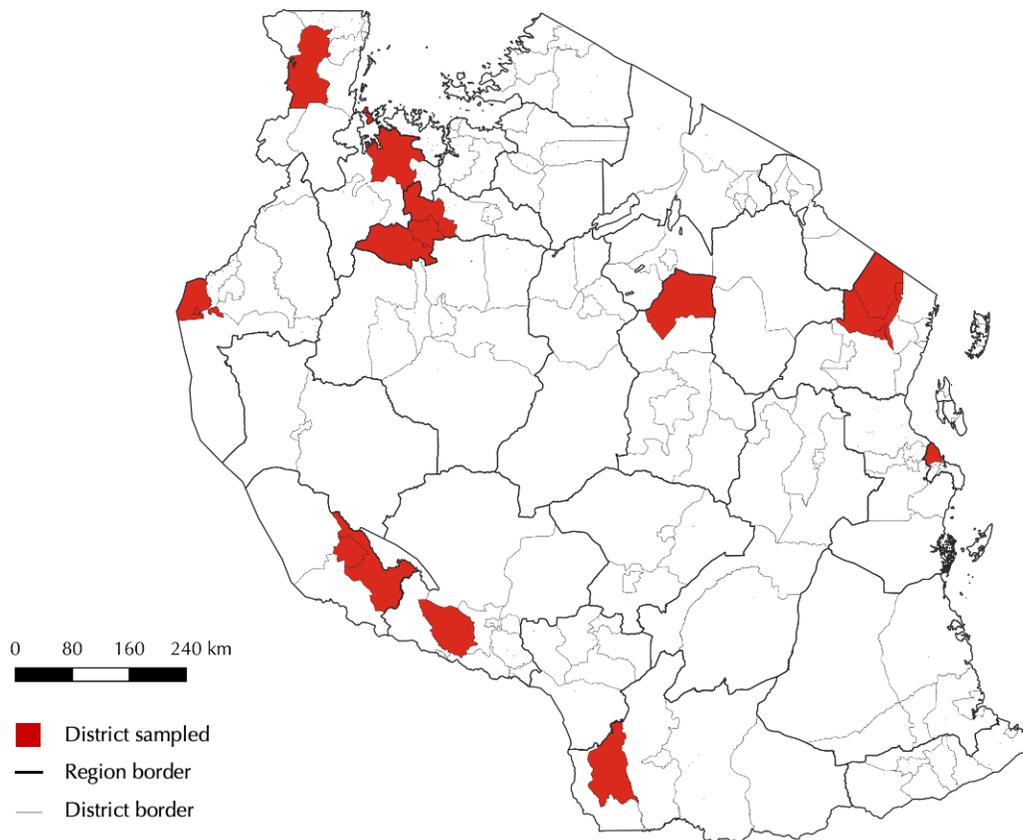
Appendix Figure 7: location of the ten sampled districts

d. Description of other contemporary reforms

Appendix Figure 8: description of other Big Results Policies in education passed around the same time as the curricular reform

I.	Pressure to perform	
Official school ranking	Ranks all government primary (and secondary schools) by pass rates in PSLE (and CSEE). Each exam has 10 performance bands, which are classified as green, yellow, or red. Results publicly posted and widely disseminated. Both national rankings and district rankings were distributed.	Fully implemented through 2016.
School incentive scheme	Annual monetary and non-monetary incentives for primary and secondary schools that have most improved their performance in the national exams (PSLE & CSEE).	Partially implemented: 60 primary schools received financial awards in 2015. Almost 4000 non-monetary awards (certificates) distributed to primary and secondary schools starting in 2016. No awards prior to 2015.
II.	Teacher motivation	
Teacher motivation	Providing both non-monetary incentives (certificates) to high performing teachers, as well as clearing all outstanding payment arrears for teachers.	Partially implemented: outstanding claims reduced by a third by 2016.
III.	Back to basics	
National 3R assessment	Early learning assessments (Grade 2) under the 3R (reading, writing, and arithmetic) assessment program on a set of randomly selected schools.	Implemented: assessment conducted in 2016.
3R Teaching Training	Teacher training program for Grade 1 and 2 teachers on how to teach reading, writing and arithmetic most effectively to this age group. Through a cascade model 37.5% of schools in 40 low-performing districts (out of 136) will be trained.	Implemented: Almost 60,000 teachers were trained by 2016. Training started in 2014.
Student Teacher English Programme (STEP)	STEP trained primary and secondary school teachers on how to identify and support low performing students. Teachers were trained on how to conduct diagnostic tests to determine which students need extra coaching, as well as how to develop curriculum and conduct classes for low performing students.	Partially implemented: teachers from 5500 primary schools were trained. Training started in 2014.
IV.	School management and finance	
School improvement toolkit	The programs aims to train head-teachers of primary and secondary schools on best practices in the management of schools. A practical toolkit of these practices was distributed to head-teachers.	Implemented: More than 16,000 primary schools received the materials. Distribution started in 2014.
Capitation grants	Timely disbursement of sufficient capitation grants for primary and secondary schools; as well as equalization of funding per student per district (about USD 4.6 per primary student and USD 11.6 per secondary student).	Late implementation: prior to 2016, 31% of schools received funds on time. In 2016 about 90% of schools received funds on time.

Notes: The delays in implementation were due to lack of funding. BRN was not adequately funded until donors provided funding in 2015. The information displayed in this table is compiled from a series of World Bank Project Implementation Status Reports and Tanzania Government reports.

e. Sample test booklets

Math Test – Grade 1

2013 Test		2014 Test	
TAJA IDADI	**UTAMBUZI WA NAMBA**	**TAJA IDADI**	**UTAMBUZI WA NAMBA**
□ = ♥♥♥♥	23 38 44	□ = ✿✿✿✿	7 58 63
□ = ♥♥♥♥♥♥♥	60 85 69		
□ = ♥		□ = ✿✿✿✿✿✿✿✿✿	**NAMBA IPI NI KUBWA**
□ = ♥♥♥	**NAMBA IPI NI KUBWA ZAIDI**		9 au 6 70 au 46
□ = ♥♥♥♥♥♥♥	13 au 24 94 au 23	□ = ✿✿✿✿✿	25 au 57 67 au 76
□ = ♥♥♥♥♥	87 au 56 44 au 63		
	91 au 19 57 au 23		

2015 Test		2016 Test	
TAJA IDADI (ID)	**UTAMBUZI WA NAMBA (UTA)**	**TAJA IDADI (ID)**	**UTAMBUZI WA NAMBA (UTA)**
□ = ★★★	5		6
	37	□ = ★★★★	46
□ = ★★★★★★★★	82		93
	NAMBA IPI NI KUBWA (BWA)	□ = ★★★★★★★★★★	**NAMBA IPI NI KUBWA (BWA)**
□ = ★★★★★★★★★★★	7 au 5		6 au 4
	19 au 27		18 au 26
	92 au 69	□ = ★★★★★★★★★★★★★★	97 au 79
	203 au 100		206 au 100

Math Test – Grade 3

2013 Test		2014 Test	
KUJUMLISHA	**KUTOA**	**KUJUMLISHA**	**KUTOA**
67 454	169 454	5 525	6 784
+ 34 + 34	- 25 - 134	+ 4 + 258	- 5 - 326
303 261	783 468	9 906	7 674
+ 259 + 130	- 232 - 135	+ 5 + 48	- 2 - 414
367 314	754 870	43 269	69 823
+ 616 + 269	- 229 - 254	+ 32 + 410	- 34 - 270
		59	72
		+ 27	- 5

2015 Test		2016 Test	
KUJUMLISHA (J)	**KUTOA (T)**	**KUJUMLISHA (J)**	**KUTOA (T)**
4 + 1 = □	6 - 5 = □ 148	5 + 1 = □	5 – 4 = □
	- 24		
12 46	11 86	11 15	11 26
+ 3 + 39	- 3 - 24	+ 4 + 5	– 4 – 14
15 213	25 327	187 36	137 97
+ 5 + 201	- 13 - 162	+ 74 + 29	– 25 – 35
166 3015	864	211 4027	314 773
+ 84 + 280	- 707	+ 203 + 190	– 161 – 608

Swahili Test – Grade 1

2013 Test

SILABI

ba	he
fu	da
sa	li
me	no

MANENO

panga	soma
kula	dada
pika	safi
pete	picha

SENTENSI

1. Damasi anasoma kitabu.
2. Paka anawinda panya.
3. Baba anakata mti.
4. Mama anapika wali.
5. Taa inawaka usiku.
6. Kaka anacheza.

KUANDIKA MANENO - IMLA

kalamu, meza, papai,
dawa, redio

2014 Test

SILABI

mi	se
ra	po
fo	

MANENO

kamati	duka
sifuri	chuma
kiazi	chandarua

SENTENSI

1. Mama amepika chakula kitamu.
2. Kabati imejaa vitu.
3. Watoto wameingia darasani.
4. Wanafunzi wameimba vizuri.
5. Nitakuja sasa hivi.
6. Ukuta wa shule umejengwa.

2015 Test

SILABI

da   mwi   njwa

MANENO

rula	dagaa
bomba	paka
kamasi	chenza

SENTENSI

1. Mtoto anasoma kitabu.
2. Juma analima shamba kubwa.
3. Shule yetu imepata walimu wapya.
4. Dereva anapiga honi ya gari.
5. Baba mdogo ametuletea zawadi.
6. Mwanafunzi amepoteza kitabu na kalamu.

2016 Test

SILABI

mo   chu   nywe

MANENO

paka	sifuri
lengo	kuni
sadaka	ngedere

SENTENSI

1. Bustani ina maua.
2. Kitabu hiki kina hadithi nzuri.
3. Wazazi wamenunua makochi mapya.
4. Kaka anakata mti.
5. Kuta za nyumba zina nyufa nyingi.
6. Tembo ni mnyama mkubwa wa porini.

Swahili Test – Grade 3

2013 Test

AYA

Zahanati ya kijiji chetu ina jengo moja. Daktari wetu anaitwa Juma. Yeye anawapenda watu na ana huruma. Juzi ametoa ushauri kuwa tutumie vyandarua ili kuzuia tusiumwe na mbu.

HADITHI

Mwaka huu wanakijiji wameamua kutumia utaratibu wa kilimo cha pamoja. Wakati wa masika waliandaa mashamba. Mvua za kwanza ziliponyesha walipanda mbegu. Baada ya siku chache mimea ilichipua.

Ilipofika wakati wa palizi waliweza kufanya kazi hiyo kwa ufanisi. Walipalilia kwa kushirikiana bila kutegeana miongoni mwao. Jitihada hizi ziliwafanya wanakijiji waweze kulima eneo kubwa na wakapata mavuno mengi.

Maswali

1. Wanakijiji waliamua kutumia utaratibu gani ya kilimo?
2. Ni wakati gani wanakijiji huandaa mashamba yao?
3. Kwa nini wanakijiji waliamua kushirikiana?

2014 Test

AYA

Dada amenda kwa shangazi. Shangazi anaishi Mnazini. Dada amefurahi kufika huko. Ameonana na binamu zake. Wote wamefurahi kumpokea. Wameandaa chakula na vinywaji.

HADITHI

Miaka mingi iliyopita kijiji chetu hakikuwa na maji ya bomba. Mwaka huu kijiji kimepata maji. Wanakijiji walishirikiana kuchimba mferejii wa kupitisha maboomba. Mafundi walifanya kazi ya kuunganisha maboomba ndipo kijiji kikapata maji.

Siku ya kuzindua mradi wa maji wanakijiji walikusanyika. Kila mmoja alitaka kuona maboomba yakitoa maji. Mgeni rasmi alitoa hotuba fupi na kufungua bomba la maji. Maji yalitiririka kwa wingi toka bombani. Wanakijiji walisherehekea kwa kuimba nyimbo na kucheza ngoma.

Maswali ya ufahamu

1. Wanakijiji walishirikiana kufanya nini?
2. Wanakijiji walisherehekea nini?
3. Kwa nini wanakijiji walifurahi?

2015 Test

AYA

Leo shule yetu wamefika wageni. Wageni hao ni wanafunzi kutoka shule iliyo jirani na shule yetu. Tuliwapokea kwa shangwe na walifurahia mapokezi yetu. Ilipofika jioni, timu za shule hizi zilicheza mpira wa miguu. Shule yetu ilishinda kwa kufunga magoli mawili kwa goli moja. Juma lijalo kutakuwa na mechi ya marudiano.

UFAHAMU

1. Shuleni kwetu walifika wageni gani?
2. Juma lijalo kutakuwa na nini?

HADITHI

Tumeelezwa kuwa misitu ni mali. Hivyo lazima tuitunze. Misitu husaidia kuhifadhi udongo. Tunashauriwa tusichome misitu ovyo. Wanyama na ndege wanaoishi misituni watakosa mahali pa kuishi. Kusipokuwa na misitu nchi itageuka kuwa jangwa.

UFAHAMU

1. Misitu ina faida gani kwa wanyama?
2. Misitu ina faida gani kwa udongo?
3. Kuna hasara gani kama tukichoma misitu?

2016 Test

AYA

Leo shule yetu wamefika wageni. Wageni hao ni wanafunzi kutoka shule iliyo jirani na shule yetu. Tuliwapokea kwa shangwe na walifurahia mapokezi yetu. Ilipofika jioni, timu za shule hizi zilicheza mpira wa miguu. Shule yetu ilishinda kwa kufunga magoli mawili kwa goli moja. Juma lijalo kutakuwa na mechi ya marudiano.

UFAHAMU

1. Shule ipi ilishinda?

HADITHI

Siku moja, Amiri aliamka mapema na akaenda kutembea. Siku hiyo kulikuwa na mnada wa mifugo kijiji cha jirani. Akaamua kuwa ni vema akaangalie jinsi mnada unavyoendeshwa. Kwa kuwa ilikuwa mara yake ya kwanza kwenda mnadani, alijifunza mambo mengi pamoja na jinsi biashara ya mifugo inavyoendeshwa.

UFAHAMU

1. Kwa nini Amiri aliamka mapema?
2. Katika kijiji cha jirani kulikuwa na tukio gani?
3. Amiri alijifunza nini?